Scaling-out Ethernet for the Data Center

Yaron Haviv [#1], Marina Lipshteyn [#2], Or Gerlitz [#3]
[#] *Voltaire, Israel*
[1] yaronh@voltaire.com
[2]marinal@voltaire.com
[3]ogerlitz@voltaire.com

***Abstract*— Data center architecture is constantly evolving, with major changes now emerging in several key areas:**

- **Data center consolidation: Building larger shared (private or public) data centers instead of many smaller ones**
- **Focus on application and business services: Moving away from manual IT processes**
- **Virtualization anywhere: Servers, I/O, storage, networks, and applications are virtualized and decoupled from physical hardware**
- **Fabric convergence: Networking, storage, and inter-process communication (IPC) from multiple applications traveling over the same physical wire**

These trends have significant impact on the fabric architecture of the data center. Fabrics must now support larger-scale Layer 2 (L2) networks since server virtualization and mobility, new storage protocols, and low-latency messaging must reside on the same L2 domain. The new data center infrastructure must also be set to overcome the management complexity and address virtualization from the ground up.

The Data Center Bridging Group of the IEEE is working to enhance Ethernet to support the architectural trends outlined above. In doing so, they are replicating many capabilities already available in InfiniBand, such as class isolation, low-latency, I/O and switch virtualization, lossless traffic flows, congestion control, multi-path L2 routing, and L2 discovery and capability exchange. These new technologies are referred to as Converged Enhanced Ethernet (CEE) or Data Center Ethernet (DCE).

This document describes the challenges inherent in existing Ethernet solutions and how new scale-out Ethernet architecture can effectively address those challenges.

I. Data Center Network Scalability Challenges

A. The Need for Scalable Data Center Networks

In recent years, we have seen computation infiltrating all aspects of our lives. More content and services are digitized, requiring more storage and associated computation and network resources. As a result, data center capacities are constantly growing at a fast pace while budgets and available power remain at the same levels.

To further increase data center efficiencies, organizations are trying to leverage economies of scale. Rather than hosting multiple smaller data centers, organizations are choosing to consolidate to fewer locations each at a much larger scale. In some cases, organizations are looking to public cloud providers that can host multiple virtual data centers in the same physical location. As a result, public and private clouds are expected to grow at unprecedented rates.

One way to enable large-scale data centers is to increase server densities. New designs allow nearly 100 nodes per rack, and more than 1000 CPU cores per rack. However, these designs require a greater emphasis on power and cooling and create new challenges in switch cabling and switch densities. As an example, 1-2 server racks may have more connections than some of the largest 10GbE switches in the market.

The key technologies enabling increased data center efficiency are virtualization and automation solutions that can squeeze more virtual servers into the same physical resources. These solutions also automate many day-to-day tasks such as delivering a new computation service, conducting maintenance and migrating loads among different hardware platforms.

As data centers grow and become denser, more virtualized and automated, the load over the underlying fabric increases significantly. Today we see the following trends emerging:

- Extensive use of shared external file (NAS) or block (SAN) storage drives significant amounts of traffic throughout the data center, with increasing demands for high reliability, high service quality, and high peak performance.
- Server virtualization forces much higher capacities over the same physical node, requiring equivalent capacity increases in the attached NICs and networks.
- Server and application mobility drives more communication between different physical segments/racks, and server migration—which requires moving the entire virtual server memory footprint from one node to the other—adds to the load.

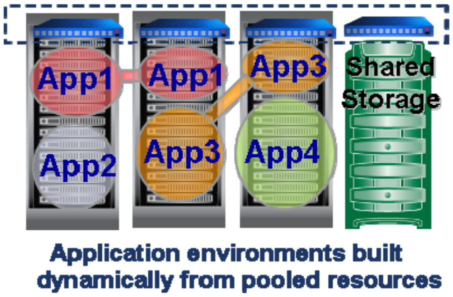
App1
App1
App3
Shared Storage
App2
App3
App4
Application environments built dynamically from pooled resources

Fig. 1: Next generation data center architecture

With the need for greater efficiency, clustering or scale-out technologies are becoming more widely used. Examples of multiple servers that are interconnected and deliver the same logical function can be found in web clusters, a database clusters, clustered file systems or Map-Reduce processing in cloud computing. Scale-out and application clustering technologies in these environments make extensive use of messaging and data movement/replication. This not only increases the load on the fabric, but also requires lower latencies, lossless and predictable behavior, and high burst performance.

It is important to note that much of the aforementioned traffic is limited to the same L2 (bridge) network domain—a virtual machine migrates with its IP address and cannot be mobilized to another IP network segment. In addition, some storage or messaging protocols (such as FCoE, RDMAoE, and PXE) are limited to L2. It is clear that new data center networks need to support much larger L2 switching domains, and cannot use L3 routing for out-of-the-rack communication as they could before.

To summarize, next generation data center networks require:

- A very large number of nodes at higher densities
- Lower power consumption
- Higher bandwidth per port
- Less bandwidth aggregation between tiers
- Lower latencies and predictable behavior
- Multiple, large, L2 (bridge) domains
- Segment isolation (partitioning) and traffic class isolation (CoS)
- Virtualization and virtualization fabric awareness

Apart from the technical requirements, all of the above capabilities need to carry a reasonable price tag for both operational and capital expenditures. Unfortunately, the above requirements cannot be met by most existing Ethernet products. A new category of scalable and data center-optimized switches must be developed to address such challenges.

B. Traditional Data Center Networking Architecture

In legacy data center designs, the data center was divided into physical silos, with each silo containing a set of servers or a rack that ran a specific application (in one or more application tiers). The application had little communication with the external world since most of its intensive transactional, messaging, and data/storage traffic ran within the rack and only a fraction of that traffic was delivered to consumers outside of each rack (see Figure 2).

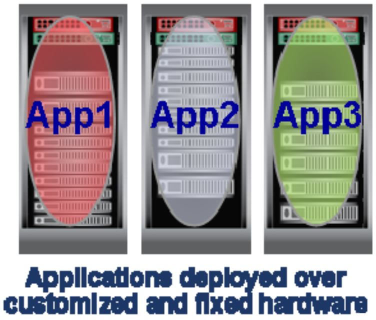

Fig. 2: Legacy data center architecture

The legacy architecture shown in Figure 3 is enabled by top of rack (TOR) access layer switches that handle internal L2 communication (bridging), with a small set of uplink ports connected to core or distribution aggregation switches, resulting in high oversubscription. In many cases the silos have a unique IP subnet, and the aggregation switches implement L3/L4 routing between those racks/silos and external consumers or other silos.

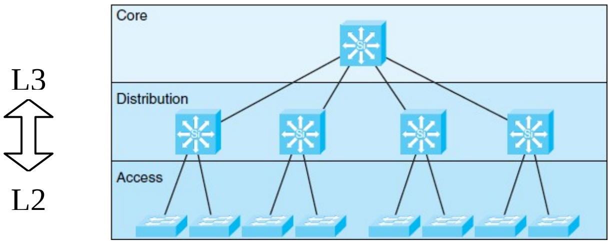

Fig. 3: Traditional 3-tiered network design

In legacy environments, core and distribution switches were always designed to support many complex network services for LAN, WAN, and enterprise communication with deep packet inspection and manipulation capabilities, large IP (L3) routing tables, large content addressable (CAM) tables, and some computation-intensive tasks like web/XML processing, encryption, and long distance optics. Because of their design and evolution, aggregation switches have a very high price per port, and much higher power consumption than access or blade switches. The table below shows the significant differences in price, power, latency, and density between L2+ TOR/Blade switches and L3+ core switches.

TABLE I
L2+ TOR/Blade switches vs. L3+ core switches

Platform	Power/ Port	Price/ Port	Latency	Max wire speed ports
10GbE TOR/ Blade	7-10W	$400-900	0.3-5 us	48 in 1U
10GbE Core	35-100W	$2600-5000	>10 us	140 in >20U
Difference	~5-10X	~5-10X	~5-10X	Only 3X ports (much less dense)

An interesting comparison to the 10GbE switch platforms described above can be made with 40 Gb/s InfiniBand switch products. These switches can deliver almost 4X the port performance at similar costs, efficiencies and power to an Ethernet access switch due to some key architectural differences between the technologies that will be elaborated later on in this paper.

TABLE II
InfiniBand core switches

Platform	Power/ Port	Price/ Port	Latency	Max wire speed ports
40 Gb/s InfiniBand core	7W	~$1000	0.3 us	324 / 648 40Gb/s ports

In Ethernet environments, there is traditionally a very high oversubscription rate (5:1 to 10:1) between server-facing ports and aggregation ports, resulting in very high costs and increased use of power in core/aggregation switches. These issues are mitigated by the fact that there are fewer ports than the number of servers in these networks. However, with the growing requirements for CPU capacity and the introduction of server virtualization, such aggregation will no longer be acceptable.

C. Difficulties In Scaling Data Center Networks

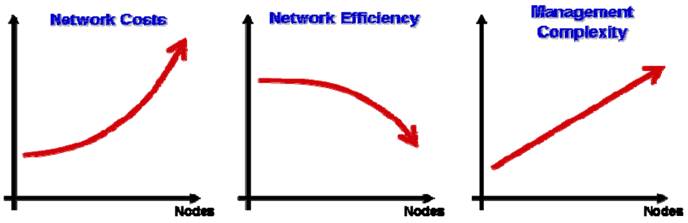

Fig. 4: Results of Ethernet’s poor scalability

Existing network switching and software solutions don’t scale well. As networks grow, the cost grows exponentially, efficiency drops, and management complexity increases. As shown in Figure 4, customers building large data centers pay much more but get much less per each additional capacity growth unit.

1.) Cost implications: As described in the previous section, traditional networks are designed with lower cost blade or TOR (top of rack) switches interconnected by much more expensive (and power hungry) aggregation switches. As the size of networks increase, additional aggregation tiers are needed to connect smaller network segments together. As a result, for every usable server node, multiple network ports are used to connect switches. The ratio of server ports to network ports increases with scale, and as we scale, we use more of the expensive (aggregation) switch ports.

Until recently, users mitigated the cost and scalability problem by using a very high oversubscription ratio between aggregation tiers. As an example, a 48 port switch was connected to 32-36 servers and had only four uplinks to the core switch. In this case a 256 port 1 GbE core switch could support 2,000 connected servers. Anything beyond this number of servers would require another switch tier.

However as described previously, changes in the data center are driving lower oversubscription rates as well as a transition from 1 GbE to 10GbE. This results in much higher network costs in large-scale environments.

2.) Performance degradation: Current network designs are not only expensive in large scale, but are also less efficient. The oversubscribed, hierarchical nature of the network creates bottlenecks as we leave the rack communication path between nodes on different racks traversing multiple aggregation switches. This adds significant latency (especially since aggregation switches are slow) and exposes the communication to network congestion (which can easily occur due to the high oversubscription ratios). When congestion occurs in the network, switches drop packets to notify the source about that congestion, which only causes greater delays that impact application performance.

The traditional Ethernet bridging protocols (spanning-tree) are designed to avoid loops, even at the expense of performance degradation. If the network contains multiple possible paths between end-points, the protocol disables all of the ports that may lead to the same destination. This behavior significantly affects network scalability since the overall network bandwidth (bisectional bandwidth) is limited to the bandwidth supported by the root aggregation switch.

In fabrics such as InfiniBand and Fibre Channel, multiple paths are allowed and managed by a fabric manager, which maximizes the utilization of all ports in the fabric. While guaranteeing a loop-free fabric, multiple root switches can run in parallel (linearly scaling the bandwidth), or different mesh topologies can be applied. The IEEE and IETF organizations are currently defining extensions to Ethernet that will allow similar behavior for Ethernet.

3.) Device-oriented system management: Most network management applications are built using a device-centric approach. Each device exposes some standard or proprietary APIs/MIBs, and the management station identifies the API/MIB as a managed object and conducts its operations at the device level.

Typical management systems show network topologies (interconnected devices), aggregate events and alarms from multiple devices, and allow some device-level configuration. In this case, routers or firewalls at junction points enforce the networking policies.

However, today's device-oriented management systems are not suitable for tomorrow's data center, which requires automation and virtualization to deliver the infrastructure as a service. Moreover, there is a disconnect when it comes to translating application-level requirements into fabric policy and configuration and a very low degree of automation. Another challenge is that network provisioning tools are implemented separately from monitoring tools. As a result, it is quite difficult to track the impact of manual or automatic policy changes on the network and application behavior.

As networks grow, the complexity of managing them grows proportionally. Future management solutions must focus on the services delivered by the network (such as connectivity, security/isolation, QoS/CoS, availability, and statistics) rather than on the devices forming the network. Such an approach can deliver much greater scalability and can more effectively address data center automation and virtualization requirements.

II. Scale-out Data Center Fabric Architecture

New data center fabric architectures can be designed to address the new challenges IT organizations face, such as:

- Virtualization and virtualization fabric awareness
- Data center consolidation
- Extensive use of server virtualization
- High density racks and the use of powerful CPUs
- Automation and cloud computing service orientation

The following sections will describe such new scale-out fabric architecture based on Converged Enhanced Ethernet (CEE), which can deliver a scalable and efficient fabric for servers and storage in the data center.

A. Efficient Cost, Power, Latency & Density

CEE fabric solutions can be designed around very high-density switches that switch many 10 GbE ports at wire speeds and without oversubscription. In addition, the switch architecture can provide several traffic management capabilities that reduce the latency to less than one microsecond, while at the same time guaranteeing application performance and lossless or lossy behavior for the various traffic classes.

The scale-out design of the switch couples less complex and less expensive hardware elements with a scale-out software stack, enabling lower costs, lower power consumption, and higher switching density than other aggregation switches available today. Furthermore, multiple switches can be meshed together to form enormous topologies without losing these advantages.

B. Linear Scalability

New IETF and IEEE enhancements (such as the TRILL protocol) would allow mesh configurations based on simpler building blocks. Some intermediate vendor enhancements can be implemented with Ethernet switches to address scale-out prior to full standard availability.

This model is unprecedented in a market where the typical solution has an exponential cost model and demonstrates server performance degradation when trying to scale-out an environment.

Figure 5 compares a solution using an off-the-shelf top of rack switch (TOR) connected to scale-out fabrics serving as an aggregation layer with the alternative, which uses traditional Ethernet aggregation switches in a hierarchical design. This clearly demonstrates the advantages in cost and efficiency in larger scale configurations.

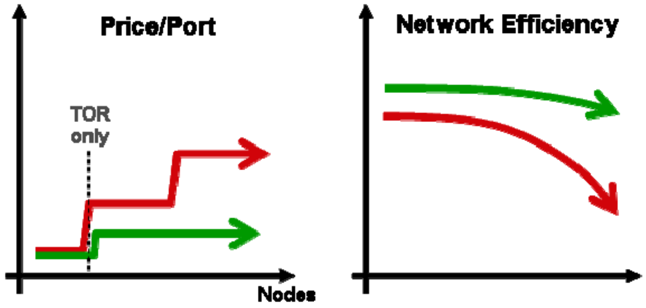

Fig. 5: Scale-out solution scalability (green) vs. other networking solutions (red)

Thus, scale-out solutions allow users to build larger consolidated server and storage farms at much lower costs while maintaining the highest performance levels.

C. Fabric Wide Virtualization

Scale-out fabrics require a fabric manager (FM) that will aggregate the management of the different fabric elements. The FM discovers all of the physical switches and virtual switches on the network, and dynamically configures them to deliver on the desired service or application requirements. The FM constantly monitors the switches and the traffic flowing through them to ensure a properly managed network and adherence to the user policies.

The FM accepts application level requirements or definitions of virtual entities and constantly adjusts the end-to-end fabric policy across all switches to accommodate the requirements, eliminating the need to manually configure the individual switches, track the physical to virtual relations, or understand how the switches are interconnected.

One of the key services delivered by a switching infrastructure is to connect multiple end-points and allow traffic to flow between them, as well as between the end-points and external connections/ports. The FM follows this concept by defining and managing virtual or physical end-points, and enforcing the traffic policy between them, regardless of their physical locations. The FM can also segregate a single physical end-point to multiple virtual end-points (in case multiple virtual machines or VNICs reside on the same node).

Examples of managed end-points include:

- A NIC, HCA or HBA on a virtual or physical machine (Virtual I/O)
- Storage elements (target, LUN/Volume, or file server)
- Router/gateway port or uplink port

Multiple end-points can be grouped as well, such as when a set of end-points in a virtual network needs to communicate on the same L2 domain and when an application cluster contains a set of network interfaces—one per every node in that cluster.

As data centers adopt converged fabrics, it is critical to define the specific class of a virtual I/O end-point. This will determine its default policy and behavior. The three main types of virtual I/O end-points are: 1.) Network adapter (NIC), with traditional LAN characteristics, 2.) Storage adapter (HBA) (requires lossless fabric behavior), and 3.) Messaging/IPC adapter (HCA) (requires low-latency and lossless behavior).

Virtual end-points may change their physical locations dynamically, such as when a virtual machine migrates from one node to another. In this case, the policy that describes the end-point behavior, or the connectivity between any two or more end-points, is maintained, and traffic monitoring is kept in sync.

D. Focus on Application Performance and SLA

The FM should allow users to define applications, application I/O and network requirements, and application flow requirements (connections between application entities). The intelligent resource manager in the FM should act as an optimizer, and automatically produces the traffic policy for all switches to guarantee the application behavior. It also maximizes performance and reports resource conflicts to users or external automation tools.

As an example, an application may have a few tiers connected in a certain topology. It may also have certain requirements for storage traffic and external (uplink) traffic, and may require low-latency for its inter-messaging communication. These requirements are easily modeled and stored using an FM. When an application is started and physical server and storage resources are assigned to these applications, the FM will configure the switching infrastructure to provision the desired topology by partitioning the fabric and creating virtual I/O end-points, and by enforcing the traffic between the end-points according to the specified policy.

The FM constantly samples the traffic on the virtual I/O entities and application flows, reporting bottlenecks or statistics back and mapping to the application objects. A user may then decide to change his/her preferences to improve application performance.

An application-driven fabric resource management and monitoring FM enables increased fabric utilization, increased application performance, isolation of applications, minimized cross interference, and much greater visibility into the application performance.

E. Service Oriented Management

As illustrated earlier in this paper, scale-out fabric solutions focus on the services delivered by the fabric, which can consist of hundreds of interconnected switches. This is achieved through application and service modeling, and through application-oriented monitoring.

As illustrated in Figure 6, the fabric is mapped to three layers, including physical objects (such as servers, switches, storage, and ports), virtual objects (such as virtual machines, virtual I/O, and volumes), and application objects (such as application tiers and connections). In addition, there are physical or logical group objects that define a collection of physical or logical individual objects. A rack, for example, is a group of servers in the same location. A user can monitor, read, or modify the objects and their attributes, or even get notified of status changes or thresholds associated with those objects.

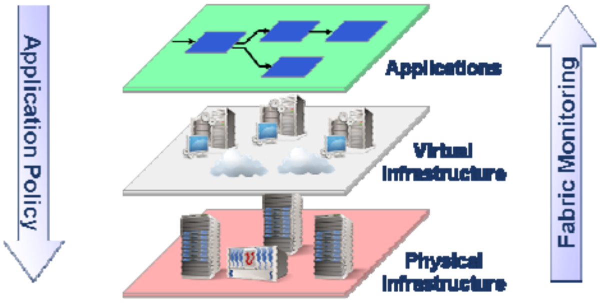

Fig. 6: Three-layered data center model

Users or external automation tools do not need to manage individual devices and their attributes. Instead, they can define the desired fabric services and allow the fabric manager to enforce these services and provide high-level feedback. The fabric can be easily managed through an extensible web-services API, through an object-oriented context aware CLI, or through a web-based graphical interface.

III. Delivering Fabric as a Service (FaaS™)

A. The Need for FaaS™

As more and more IT organizations focus on consolidating data centers, simplifying their operations, and implementing automation concepts, they can pool server and storage resources and assign them to applications on demand.

New automation and virtualization technologies allow treating Infrastructure as a Service (IaaS), and brokering user requests with available resources. This is further enhanced with private or public cloud architectures that allow "renting" of server infrastructure or even complete applications.

Data center consolidation cannot be achieved without controlling the underlying fabric, and since the environments are automated and managed through service concepts, one cannot manually configure the fabric. Thus, a service-oriented fabric management paradigm is needed.

The key challenges in data center consolidation and cloud computing that relate to the fabric are:

- There is a lack of isolation and security between virtual data centers
- There is a lack of service level monitoring and enforcement for shared fabrics
- Virtualization for servers and storage have a weak correlation to fabric policies
- Application mobility and virtual machine (VM) migration require synchronization with fabric policies
- With CPU consolidation and virtualization bottleneck transfers to I/O, I/O and fabric resource optimization is needed to avoid bottlenecks and ensure application performance
- The placement of jobs and VMs over the fabric while taking into consideration the fabric layout, current load, and application I/O requirements is challenging
- Scale-out environments are challenging for administration and troubleshooting
- Measuring the impact of fabric congestion or oversubscription on application performance is difficult

The above challenges are addressed by the Fabric as a Service (FAAS) approach, which extracts server interconnect networks from their physical elements and controls them as variable logical entities.

Key benefits of the FaaS architecture include: 1.) Delivering application level SLA, performance monitoring and optimization, 2.) Controlling multi-tenant and virtualized applications in a single large fabric, and 3.) Enabling complete data center automation.

B. Fabric Manager (FM): Delivering FaaS

An FM delivers FaaS by pooling and owning all of the fabric resources—including virtual or physical switching elements and I/O adapters—and by providing central fabric monitoring and service-oriented fabric policy enforcement.

Services delivered by the FM can include the following:

- Collect statistics and information from physical and virtual switches
- Generate statistics and traffic analysis per VM, specific I/O, or traffic flow
- Carve the fabric into multiple classes of service for LAN, IPC, and storage
- Apply QoS (priority, limits, guarantee) per I/O or traffic flow
- Physical partitioning, VLAN, and virtual I/O provisioning
- Guarantee HA (multi-rail, multi-path configuration, and policy synchronization)
- Centrally manage multiple switches through a single console
- Suggest optimal placement based on fabric allocation or load
- Congestion isolation, control/throttling, and monitoring

The fabric management life cycle can consist of the following:

1. Application requirements are characterized by users or external orchestration or automation tools via the GUI, CLI, or web-services API.
2. Physical or virtual resources are assigned to the application templates manually or automatically via an external scheduling or resource management tool.
3. Fabric is configured and optimized to deliver on the desired application policy and maximize application performance.
4. Statistics and status information is gathered from all switching elements and optional agents, and mapped to the applications and application flows, generating alarms if needed.
5. Statistics and fault information can be used to manually or automatically adjust fabric behavior.
6. Users or automation tools can apply changes such as migrating virtual machines, increasing or decreasing capacities, and changing connectivity to running objects, resulting in fabric re-adjustments.

IV. Summary

As data centers grow and become denser, more virtualized, and automated, the load over the underlying fabric increases significantly along with the need for greater efficiency. As a result, clustering or scale-out technologies will be more widely used. InfiniBand fabrics were designed from day one as a scale-out data center fabric with a very lightweight switching infrastructure, the ability to run mesh topologies and multiple paths, and with fabric and I/O partitioning capabilities and central discovery and policy management. However, traditional Ethernet products do not scale well. As networks grow the costs grow exponentially, efficiency drops, and management complexity is increased.

New scale-out fabric architecture based on Converged Enhanced Ethernet (CEE), can deliver a scalable and efficient fabric for servers and storage in the data center with the following advantages:

- Efficiency: The scale-out design couples less complex and less expensive hardware elements with a powerful scale-out software stack, enabling lower costs, lower power consumption, and higher switching density than other aggregation switches currently available.
- Linear scalability of latency, power and performance: When using scale-out solutions, both cost and performance are linear, so whether the topology consists of a few hundred connected nodes or a few thousand connected nodes, it will have exactly the same price per port, exactly the same latency, and, the same total-bandwidth per port.
- Virtualized from the ground up: A Fabric Manager (FM) that discovers all of the physical and virtual switches on the network, and dynamically configures them to deliver on the desired service or application requirements.
- Focus on application performance and SLA: The intelligent resource manager in the FM acts as an optimizer, and automatically produces a traffic policy for all switches on the network, guaranteeing the application behavior.
- Service-oriented management: Users define the desired fabric services and allow the fabric manager to enforce these services and provide high-level feedback.